Nontranslated polyadenylated ribonucleic acids from the protozoan parasite *Entamoeba histolytica*

Alok Bhattacharya*,†, Sudha Bhattacharya and John P. Ackers§**

*School of Life Sciences and **School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110 067, India
§Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

Protozoan parasites display a range of unusual molecular mechanisms that may be helpful for their survival in nature. Among these parasites *Entamoeba histolytica*, the causative agent of amoebiasis, is one of the most prevalent in developing countries like India. *E. histolytica* transcribes at least four different unusual transcripts, IE, Tr, *ehapt2* and UEE1, that are polyadenylated but do not have extensive open reading frames and are unlikely to code for anything other than small peptides. In this review we describe the current status of our understanding about these transcripts.

THE protozoan *E. histolytica* is the causative agent of amoebiasis, an invasive disease that mainly affects the intestine and occasionally extra-intestinal tissues such as liver and brain. Over 50 million people world-wide suffer from amoebiasis, with majority residing in developing countries such as India[1]. Of the different species of *Entamoeba*, *E. histolytica* is known to cause invasive disease[2]. Our knowledge of *E. histolytica* genome organization is still at its infancy. While the number of chromosomes and the ploidy are still being worked out, the data available so far suggest that it may have unusual genome organization with short intergenic regions[3]. A number of circular DNAs of varying size classes have been identified[4] and the structural organization of the circular DNA carrying the ribosomal RNA genes has been worked out in detail. Unlike in many other protozoan parasites, *Entamoeba* genes have consensus sequences upstream from the transcription start site which have been shown to be involved in regulation of transcription[5]. So far presence of RNA polymerase II has not been unequivocally demonstrated[6,7].

Four different classes of poly A containing RNA have been characterized in *E. histolytica* so far which do not have large open reading frames (ORFs) that can be translated into meaningful proteins. Together they constitute the most abundant poly A-containing RNA in *E. histolytica*. It is possible that the stop codons could be introduced or removed by post-transcriptional modification as in case of RNA editing[8]. However, it does not appear that any post-transcriptional modification is taking place in these transcripts as cDNA and genomic sequences of these are exactly identical. The functions of these transcripts are still unknown.

Transcript from the ribosomal DNA plasmid

The best characterized circular ribosomal DNA molecule in strains of *E. histolytica* is the rDNA plasmid found in *E. histolytica* strain HM-1:IMSS (EhR1); this 24.5 kb molecule has been fully sequenced[9]. Each molecule of EhR1 contains two rRNA transcription units arranged as inverted repeats separated by a 3.7 kb downstream intergenic spacer and 9.2 kb upstream spacer. Both upstream and downstream regions contain different classes of repeat sequences apart from a sequence coding for a polyadenylated 0.7 kb RNA that is transcribed from the upstream region. The RNA was detected by Northern hybridization using probes from different parts of EhR1 and using RNA that was fractionated over oligo-dT column. No hybridization was detected in the unbound fraction suggesting that the transcript (Tr) has a poly A tail[10]. When the nucleotide sequence was translated to look for open reading frame, all the six reading frames had abundant stop codons. Since the same stop codons used by other *E. histolytica* genes are translational stop sites, the 0.7 kb Tr can, at best, code only for short peptides. The size of the largest ORF is 117 nucleotides. We have also analysed the Tr sequence with a new gene-prediction software (GeneScan) developed by us[11]. Figure 1 shows the power spectrum of Tr, rRNA and actin. A typical gene (the sequence that codes protein, that is, only the exons) like actin gives a peak at $f = 1/3$ and peak-to-noise ratio is much larger than 4 (Figure 1 *a*). On the other hand, a sequence that does not code for a protein, for example rRNA, does not show any such peak at $f = 1/3$ (Figure 1 *c*). GeneScan of Tr reveals a pattern similar to that of rRNA (Figure 1 *b*), confirming that Tr is unlikely to be a protein coding sequence.

The possibility of producing functional RNA by editing appears unlikely since a cDNA sequence exactly matching the Tr sequence has also been reported from

†For correspondence. (email: alok@jnuniv.ernet.in)

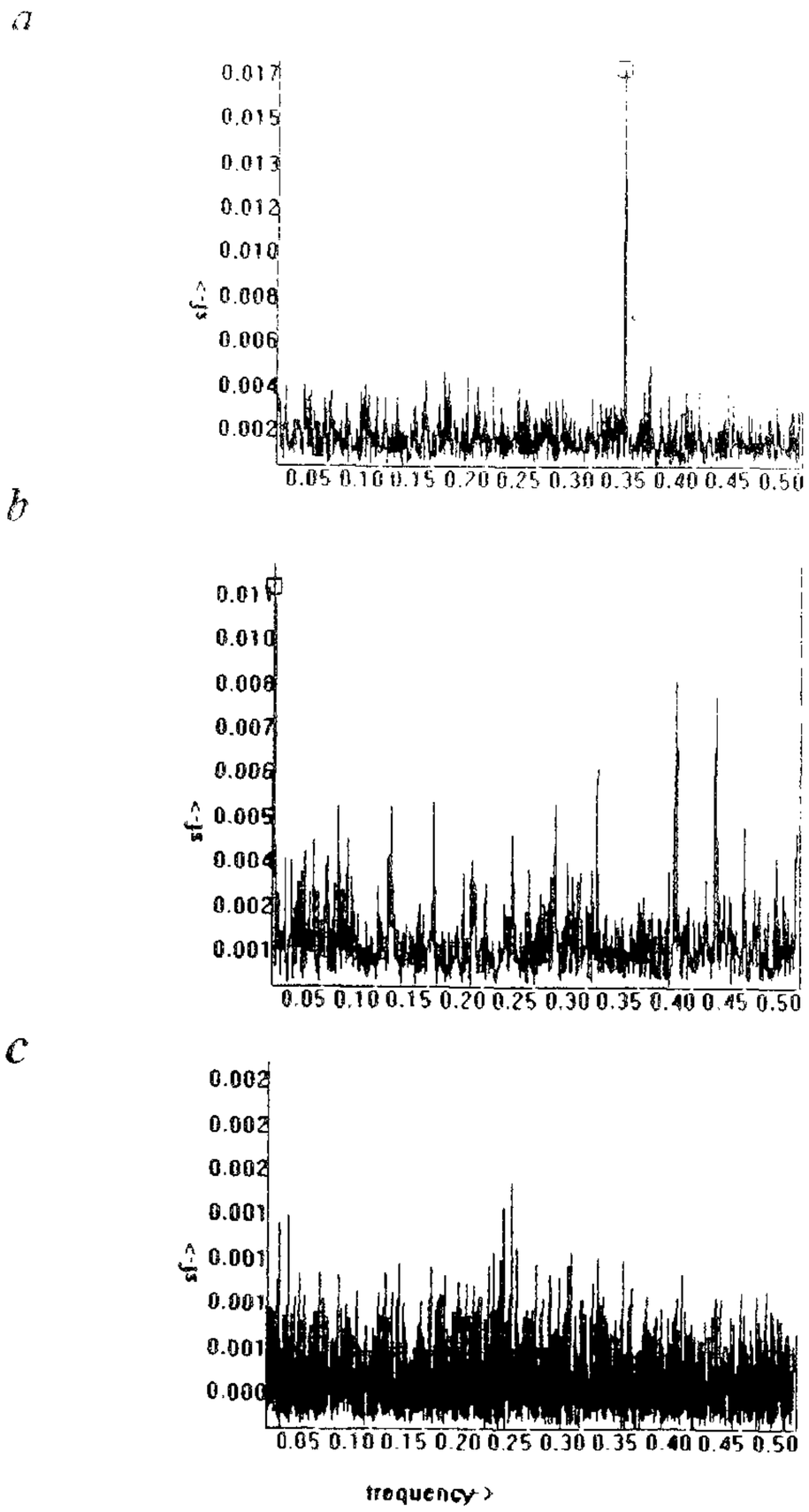

Figure 1. Analysis of coding potential. The coding potential of different sequences were analysed by the Fourier-based computational method GeneScan. *a*, *E. histolytica* actin; *b*, Tr cDNA and *c*, *E. histolytica* small subunit rRNA.

E. histolytica[12]. An interesting feature of the transcript is the presence of 26 bp repeats. There are nine of these near-perfect repeats in the Tr of *E. histolytica* strain HM-1:IMSS. The number of these repeats varies from strain to strain and it may range from eight to twenty[13]. Moreover, there are also strains of *E. histolytica* which do not have the portion of the ribosomal plasmid that encodes Tr. To date only three such strains have been identified[13]. HK-9 and Rahman are two such strains where the physical maps of the ribosomal plasmids have been determined[9]. The plasmids in these strains have only one rRNA transcription unit instead of two and the entire fragment carrying the leftward transcription unit with a portion of the upstream sequence (encoding Tr) was found to be missing in these strains. rRNA is also encoded by plasmids in other species of *Entamoeba* such as *E. invadens* and *E. moshkovskii*. In both these species there is only one rRNA transcription unit and there is no evidence for any other transcript derived from the plasmid.

The absence of an ORF of reasonable length suggests that Tr may function through its RNA. It could either have a regulatory role in transcription or translation or may have a structural role. Whatever the function may be, it is likely to involve its secondary structure. The possible folding pattern of the transcript was determined by the program RNAFOLD of the GCG (Wisconsin Genetic computer group) package and is shown in Figure 2. It is clear that Tr has a potential to fold with extensive secondary structure. The 26 bp repeat regions in the Tr exhibit maximum secondary structure within the molecule. The potential to form a folded structure of defined shape provides support to the idea that Tr may be a functional RNA.

Several eukaryotic plasmids encode *trans* acting proteins required for their stable existence in the cell[14]. Since Tr is absent in some strains of *E. histolytica*, it is

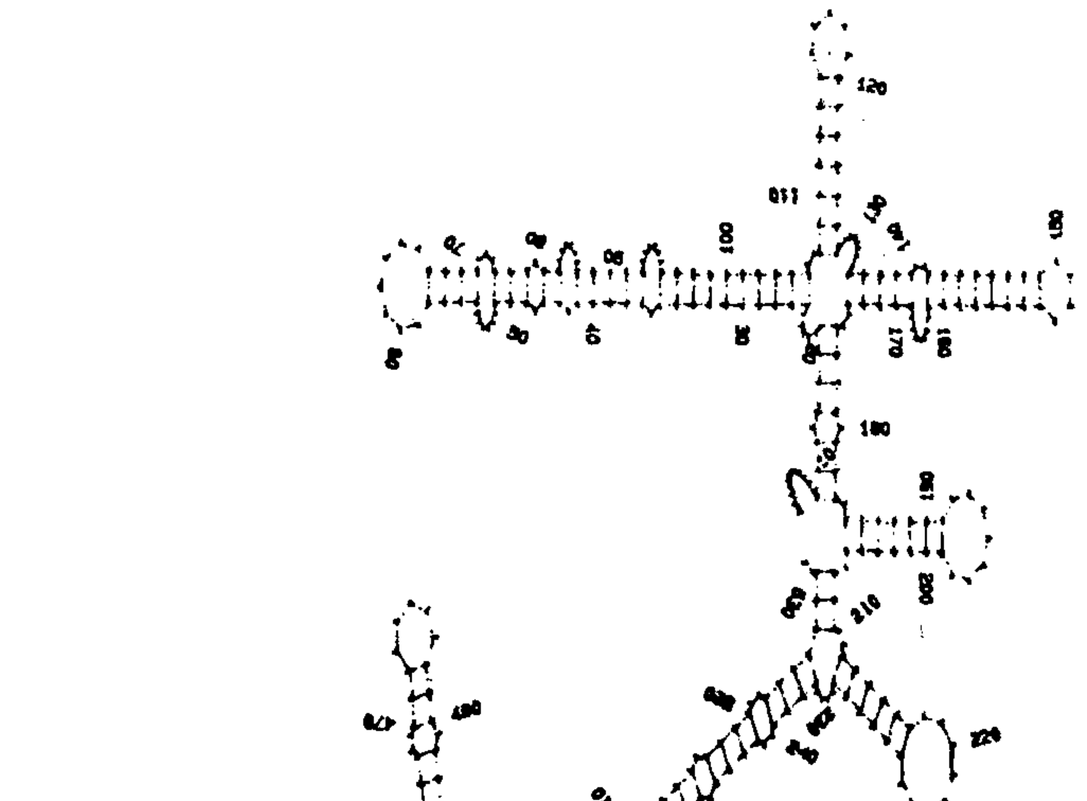

Figure 2. Secondary structure of Tr. The secondary structure of Tr was obtained by using the program RNAFOLD of the GCG package.

unlikely to be involved in plasmid maintenance. It may be speculated that it may have a regulatory role in the transcription of the leftward rRNA transcription unit. Currently there is no direct evidence in support of that.

UEE1 transcript

Our laboratory has used the 'expressed sequence tag' (EST) approach to identify novel genes from both the pathogenic *E. histolytica*[15] and the non-pathogenic *E. dispar* (manuscript in press)[16]. During these studies, a highly abundant transcript UEE1 was identified in *E. dispar* and by database search two ESTs of *E. histolytica* were also found which showed significant sequence identity with UEE1. Quantitative hybridization studies show that UEE1 is present in a number of copies (at least four per haploid genome equivalent) in the genome and represents about 8% of all the cDNAs of *E. dispar*. Not all UEE1s are identical and they display about 2–14% sequence divergence from each other. No significant ORFs can be found in any of the UEE1 sequences. The maximum length of any ORF is 40 amino acid residues. These cDNAs also have a polyadenylated tail at the 3′-end.

The IE sequence family

The first two members of this family of sequences were described seven years ago[17] and consisted of a genomic clone and a very similar but not quite identical sequence present in a cDNA library. Neither sequence contained ORFs but Northern blotting showed the presence of a 0.7 kb transcript. Apart from the axenic strain 200:NIH from which both genomic and cDNA libraries had been prepared, hybridization at high stringency showed these sequences to be present in the xenic strains TE and SI also. Under these conditions no hybridization was observed with total *E. dispar* DNA.

Subsequently, two more members of the family were described, a longer (1.2 kb) cDNA clone and another detected in a published genomic sequence. The latter is particularly interesting as it lies between two ORFs – downstream of an unknown protein and upstream of that for the *E. histolytica* pore-forming protein gene. Complex patterns from Southern blots of genomic DNA suggested that multiple copies of IE sequences were present. About 500 copies per genome are[18] estimated. IE sequences were also shown to occur in a second axenic strain of *E. histolytica* – HM-1:IMSS. Over a core region of about 500 nucleotides, all four sequences are more than 95% identical.

No further complete genomic copies have been identified, but a partial sequence is present upstream the *E. histolytica* FeSOD gene in a clone (GenBank Accession X70852) which also contains the 5′ end of an actin gene[19].

Six ESTs with homology to the IE sequences were detected in a directional cDNA library prepared from *E. histolytica* strain HM-1:IMSS[20] and a seventh (EhEST149) is also probably a member of the family. Although Tanaka *et al.* do not draw this conclusion, this means that IE sequences are the most abundant in their library. A further clone from a cDNA library derived from the *E. histolytica* strain SFL-3 has been sequenced and shown to be an IE (GenBank Accession AF126995; G. Clark, pers. commun.) and a number of others have been obtained by PCR using conserved region primers.

In all, partial or nearly complete sequences of cDNAs for about 20 RNA transcripts are now known and in no case have long ORFs been identified. The function of these abundant sequences is currently unknown. Ortiz-Garcia *et al.*, who obtained a cDNA clone (Eh61) with 71% identity to published IE sequences showed that the transcript could fold into two stem-loop structures[21]. They suggested that either IE genomic sequences or their transcripts might be involved in controlling the transcription of nearby genes, but no definite evidence to support this is available. It is also possible that the IEs represent some kind of mobile genetic element, although they do not have any obvious signature sequences to confirm this. The whole family is very unusual (no significant homology with sequences in any other organisms have been reported) and it would need further investigation to decipher the function.

The *ehapt1* and *ehapt2* transcripts

In a recent paper, Willhoeft *et al.*[22] describe the presence in a cDNA library prepared from *E. histolytica* HM-1:IMSS of two highly abundant, polyadenylated transcripts neither of which contain long ORFs. They designate these transcripts *ehapt1* and *ehapt2*. The latter comprises further members of the IE sequence family (see above) while *ehapt1* is hitherto undescribed. There is no significant sequence similarity between the two families. These authors also speculate that both transcripts might be either regulatory RNAs or be derived from transposable elements but again there is no direct evidence for either suggestion.

Conclusions

It appears that some of the most abundant polyadenylated transcripts in *E. histolytica* probably do not contain any protein coding segments. These may function as RNA molecules as suggested by presence of extensive secondary structures. So far there is no data to offer any

idea about the function of these molecules. It has been speculated that these may be involved in regulation of transcription. To our knowledge, presence of such transcripts has not been yet been reported from other protozoa.

1. Walsh, J. A., *Rev. Infect. Dis.*, 1986, **8**, 228–238.
2. Clark, C. G. and Diamond, L. S., *Mol. Biochem. Parasitol.*, 1991, **49**, 297–302.
3. Bruchhaus, I., Leippe, M., Lioutas, C. and Tannich, E., *DNA Cell Biol.*, 1993, **12**, 925–933.
4. Dhar, S. K. Roychoudhury, N., Bhattacharya, A. and Bhattacharya, S., *Mol. Biochem. Parasitol.*, 1995, **70**, 203–206.
5. Singh, U., Rogers, J. B., Mann, B. J. and Petri, W. A. Jr., *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 8812–8817.
6. Lioutas, C. and Tannich, E., *Mol. Biochem. Parasitol.*, 1995, **73**, 259–261.
7. Albach, R. A., *J. Protozool.*, 1989, **36**, 197–205.
8. Stuart, K., Allen, T. E., Kable, M. L. and Lawson, S., *Curr. Opin. Chem. Biol.*, 1997, **1**, 340–346.
9. Sehgal, D., Mittal, V., Ramachandran, S., Dhar, S. K., Bhattacharya, A. and Bhattacharya, S., *Mol. Biochem. Parasitol.*, 1994, **67**, 205–214.
10. Sehgal, D., Ph D Disertation submitted to Jawaharlal Nehru University, New Delhi, 1995.
11. Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R., *Comput. Appl. Biosci.* (currently *Bioinformatics*), 1997, **13**, 263–270.
12. Bursch, D. J., Li, E., Reed, S., Jackson, T. F. H. G. and Stanely, S. L. Jr., *J. Clin. Microbiol.*, 1991, **29**, 696–701.
13. Clark, C. G. and Diamond, L. S., *Exp. Parasitol.*, 1993, **77**, 450–455.
14. Farrar, N. A., *Trends Genet.*, 1989, **4**, 343–348.
15. Azam, A., Paul, J., Sehgal, D., Prasad, J., Bhattacharya, S. and Bhattacharya, A., *Gene*, 1996, **181**, 113–116.
16. Sharma, A., Azam, A., Bhattacharya, S. and Bhattacharya, A., *Mol. Biochem. Parasitol.*, 1999, **99**, 279–285.
17. Cruz Reyes, J. A. and Ackers, J. P., *Arch. Med. Res.*, 1992, **23**, 271–275.
18. Cruz Reyes, J., ur Rehman, T., Spice, W. M. and Ackers, J. P., *Gene*, 1995, **166**, 183–184.
19. Tannich, E., Bruchhaus, I., Walter, R. D. and Horstmann, R. D., *Mol. Biochem. Parasitol.*, 1991, **49**, 61–72.
20. Tanaka, T., Tanaka, M. and Mitsui, Y., *Biochem. Biophys. Res. Commun.*, 1997, **236**, 611–615.
21. Ortiz-Garcia, D., Meraz, M. A. and Meza, I., *Arch. Med. Res.*, 1997, **28**, 30–31.
22. Willhoeft, U., Bu, H. and Tannich, E., *Protist*, 1999, **150**, 61–70.

ACKNOWLEDGEMENTS. Some of the work presented here was carried out with financial support from DST and CSIR. We thank Dr Graham Clark for critically reading the manuscript and Dr Egbert Tannich for supplying us with unpublished results.